Exploring geolocation issues in social media analytics

A case study with Tweet messages

Daniela Carrion, Federica Migliaccio, Diana Pagliari
DICA – Geodesy and Geomatics
Politecnico di Milano
Piazza L. da Vinci, 32, 20133 Milano
daniela.carrion@polimi.it

Abstract—**Social media data, such as Tweet messages, are sometimes associated with their geolocation. This information can be exploited to perform spatial analyses, resulting in geosocial analytics. However, the geolocation does not often correspond to the actual position of the author, but could be fictiously associated to the messages. The issues coming from the absence of Tweet geolocation metadata are explored in this paper, through a test case over Italy.**

Keywords– Social media, geolocation, spatial analysis, geosocial analytics, Tweet

I. GEOLOCATING TWITTER MESSAGES

In the last years, an ever growing number of social network users have been providing a wealth of data containing a very high level of heterogeneous information. This, of course, has drawn the attention of many research areas spanning from social science and linguistics to economics and Geographical Information Science [1].

Location Based Social Networks allow to access a large quantity of volunteered semantic and geolocated contents [2]. Among these services, Twitter is the most investigated in numerous research fields such as flu spread [3], election prediction [4], earthquakes [5, 6]. The information included in the Twitter messages (Tweets) may contain not only semantic attributes (the so called "sentiment"), but also spatial and temporal data. However, the relationship between the Tweets and their geolocation could be quite complex. In fact, only about 1-3% of the posted Tweets is directly georeferenced [7], but also in this case the georeferencing could be derived from different sources and, consequently, characterized by various levels of accuracy. The Tweet coordinates encoded with the geo tag could be derived from either the IP-address, the cell-site location or the GPS module of the mobile device [8]. To our knowledge, in this case there is no way to distinguish among this kind of meta-information.

Moreover, the large majority of the Tweets does not have any "geo tag" filled, even if some information about the position can be inferred from the data stored in the user profiles or associated from the location of the Twitter places, stored in the users' Tweet list. Usually this information refers to a city, a single block or a business activity (https://support.twitter.com) and is saved in the "place" field.

Nevertheless, there is a number of studies (see, for instance, [9, 10, 11]) focused on the deduction of geographic content from the semantic information included in the Tweet text. In this case, the accuracy which can be attained in the geolocation is very limited, even of the order of hundreds of kilometers, meaning that this kind of data is useful only in case of very low-resolution applications.

The aim of the research described in this paper is to explore issues related to geolocation information provided with the Tweets, in order to investigate how their spatial analysis can give a contribution to the semantic interpretation provided by social statistics. Here the first results of this research are presented.

II. AVAILABLE DATA: TWEETS ABOUT ITALIAN SUPERMARKETS

The purpose of this paper is to evaluate the issues arising when exploiting the geolocation information of Tweet messages. The test data are Tweets regarding Italian supermarkets, e.g. messages including keywords or hash-tags concerning supermarkets. The geolocated Tweets are located in Italy and consist of 2887 messages (see Figure 1), corresponding to about 5% of all supermarket-related Tweets posted in Italy over a year and a half time span. The messages were, at first, acquired and processed to obtain statistics regarding the "sentiment" for specific supermarket chains, disregarding their location.

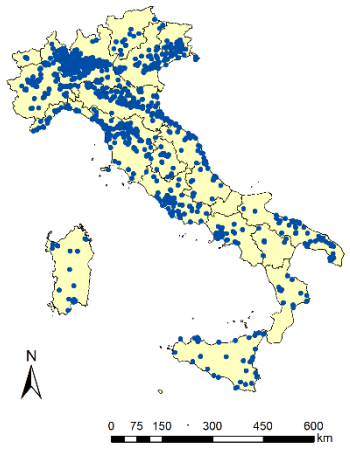

Figure 1. The geolocated Tweets of the case study

The data are in table format presenting many fields. The majority of the fields contain data on the Tweet characteristics (e.g. the message text, if it is in response to another Tweet, …)

eISSN: 1339-9144, cdISSN: 1339-231X
ISBN: 978-80-554-1325-9

and its location (Country, Continent, ...), coming either from the author's account, (i.e. the author registration), which could be invented and not real, because they are not verified, or from the data source (e.g. the Country of the server through which the Tweet was sent). The Tweet date and time stamp are reported, as well as the name of the author and information about his/her level of activity on Twitter, such as the number of followers and people who are followed. Moreover, when available, latitude and longitude coordinates are provided (as previously stated, for this test case this happens for about 5% of messages).

Unfortunately, the metadata regarding the geolocation, in terms of how latitude and longitude have been derived (GPS, IP address...), are not available. To date, it has not been possible to explore if with different data providers it could be possible to acquire more specific information regarding the geolocation.

III. Issues in Interpreting the Tweet Coordinates

The analysis of the geolocation of the available Tweet messages immediately presents issues that hinder the processing. A first spatial analysis that may be performed is to show the spatial distribution of the data points. Performing this analysis, it is evident that there are many Tweets which overlap one another. Considering the literature references [5] the reasons for this overlapping could be various, depending on the geolocation criteria as described in Section I. In Table I, the largest numbers of overlapping Tweets are shown. A significant number of Tweets overlap; the largest number of Tweets overlapping over the same coordinates correspond to 6.5% of the total. In general, 42 % of the considered Tweets overlap at least once. This kind of overlapping, with exactly equal coordinates, could be due to various reasons: e. g. people tweeting from home, i.e. the same IP address used many times, or people tweeting from a public IP address, e.g., in our case, the IP of the Wi-Fi of a supermarket. However, the reason could also be the cell-site of the mobile device.

TABLE I. Number and Percentage of Overlapping Tweets

Number of overlapping Tweets	Percentage of overlapping Tweets
187	6.5%
83	2.9%
54	1.9%
41	1.4%
34	1.2%
27	0.9%
23	0.8%
22	0.8%

The absence of metadata and the uncertainty of the nature of the Tweet geolocation affects the reliability of the positioning. A crosscheck of the other available table fields was performed to try to acquire additional information about the origin of the geolocation, such as a check with respect to the author (many Tweets from the same author, tweeting from home, would have the same coordinates), but the results of these checks could not lead to clear deductions. In Figure 2, the number of authors and the number of different posting days corresponding to the overlapping Tweets is shown. In most of cases, the number of authors, as well as the number of posting days, is lower than the number of overlapping Tweets; only in one case the numbers are the same. This means that "the same author posting from his Wi-Fi" is not the most common reason for data overlapping. In general, almost all overlapping messages have been posted in different days.

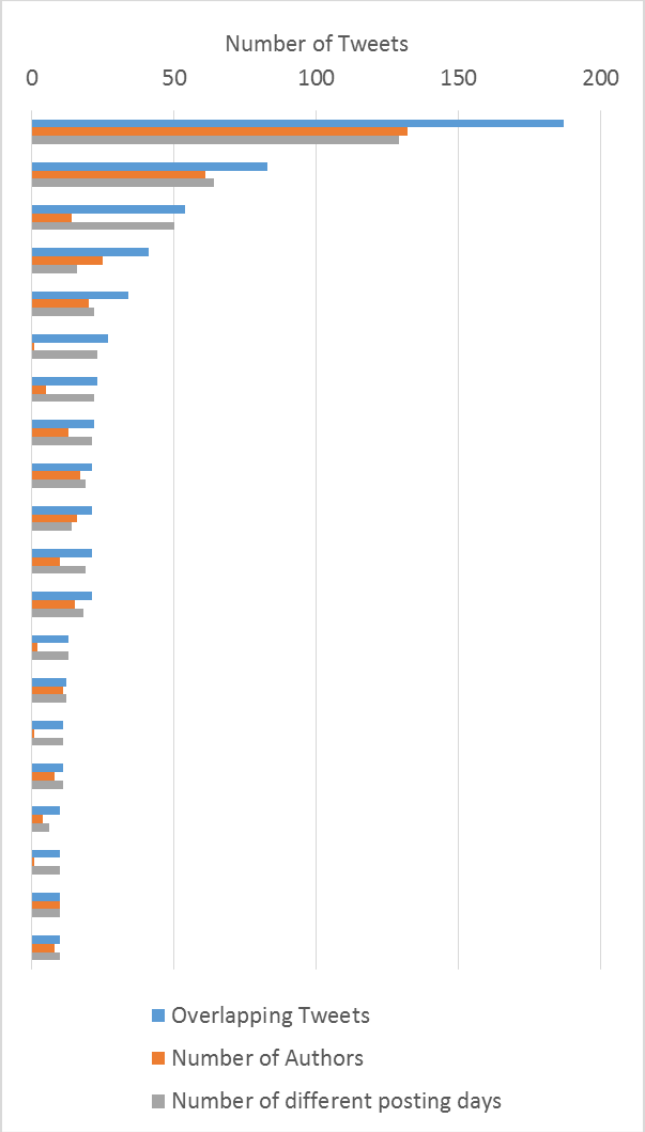

Figure 2. Number of authors and number of different posting days, corresponding to the overlapping Tweets.

IV. Examples of Spatial Analyses

Spatial analyses of data distribution have been performed to see how they can be affected by the Tweet overlapping. Since the reason of the overlapping is not known, in some cases it could affect the significance of the geolocation interpretation. For this reason, the hypothesis of disregarding the overlapping

eISSN: 1339-9144, cdISSN: 1339-231X
ISBN: 978-80-554-1325-9

messages has been considered. The standard deviational ellipses of the data falling within the Milan Municipality area have been computed and are presented in Figures 3 and 4. This is a way of measuring the dispersion of a point distribution around the mean center, also showing whether a distribution exhibits a directional trend. In Figure 3 we have represented all data, including the overlapping messages; in Figure 4 we have only represented the messages appearing in single positions.

It is evident that the act of removing the overlapping Tweets significantly affects the spatial distribution of the point features.

In order to perform a cluster analysis based on the density of Tweet messages, the number of Tweets falling into a 1 km x 1 km grid has been computed. The results are shown in Figure 5.

In the top part of Figure 5 the Getis-Ord Gi* [12] hot spot analysis is represented for all "gridded" Tweets in the Milan Municipality area. The same analysis has been repeated disregarding the overlapping Tweets and the output is shown in the bottom part of Figure 5. The presence of many overlapping Tweets in the top part of Figure 5 turns out to have a masking effect over the other data, preventing cluster identification. On the contrary, hot spots and cold spots emerge in the bottom part of Figure 5.

The conclusion that can be drawn after these preliminary analyses is that the presence of overlapping Tweets significantly affects the spatial statistics of the data. It is difficult to deduce if it is appropriate to exclude overlapping Tweets, because if the location has been critically moved with respect to the "real" one, then this could bias the results; however, the exclusion could also lead to disregard meaningful data.

V. Discussion and Conclusions

Data from social media, in some cases, come with coordinates, so they can represent a valuable source of information to detect patterns in the distribution of the mobile devices or even help to highlight trends in the pattern distribution. However, before exploiting social media data, in particular Tweet messages, for geo-social analytics, one must be aware that the positions recorded along with the text are not always corresponding to the actual position from which the text has been sent. In fact, in most cases the coordinates refer to some peculiar point or bounding box, having as a consequence the artificial accumulation of data points on a number of single locations, which may bias any subsequent spatial analysis of the data.

In this paper, an overview of the problems arising from the ways in which social media data are geo-referenced and how they are mirrored by the spatial distribution of point patterns have been presented. The case study is based on a set of over 2800 geo-referenced Tweet messages originating from Italy, which were used to mine the customer's sentiment in order to support marketing activities. Examples of spatial processing of data to highlight issues related to the absence of meta information regarding Tweet geolocation have been shown.

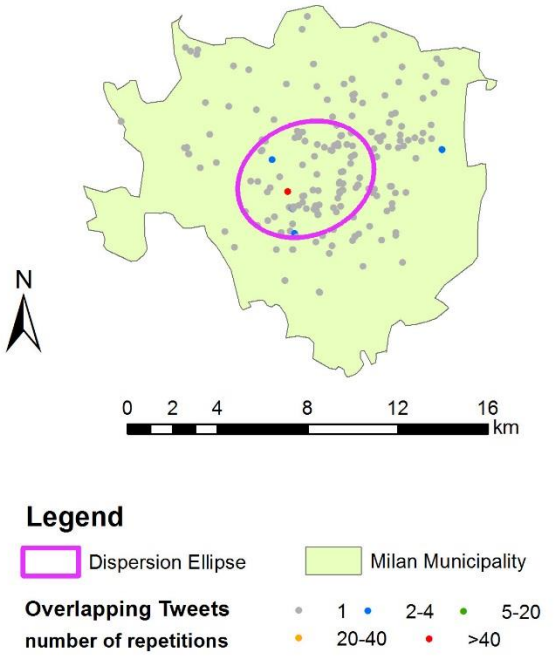

Figure 3. Standard deviational ellipse of the Tweets in the Milan Municipality area, including overlapping Tweets.

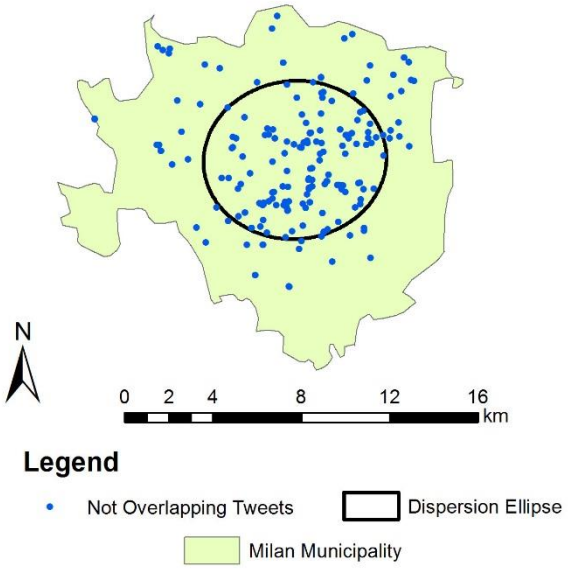

Figure 4. Standard deviational ellipse of the Tweets in the Milan Municipality area, considering only not overlapping Tweets.

One evident effect of the different ways in which Tweets are geolocated is the presence of Tweets overlapping with exactly the same coordinates. It has been shown in Section IV that the presence of overlapping Tweets significantly affects simple spatial analyses, such as the standard deviational ellipse and cluster analysis. It is difficult to infer if it is better or not to disregard the Tweets which overlap: on the one hand they for sure bias the results of spatial analysis, however, on the other

10.18638/ictic.2017.6.1.316

eISSN: 1339-9144, cdISSN: 1339-231X
ISBN: 978-80-554-1325-9

hand, excluding them could lead to ignore significant information.

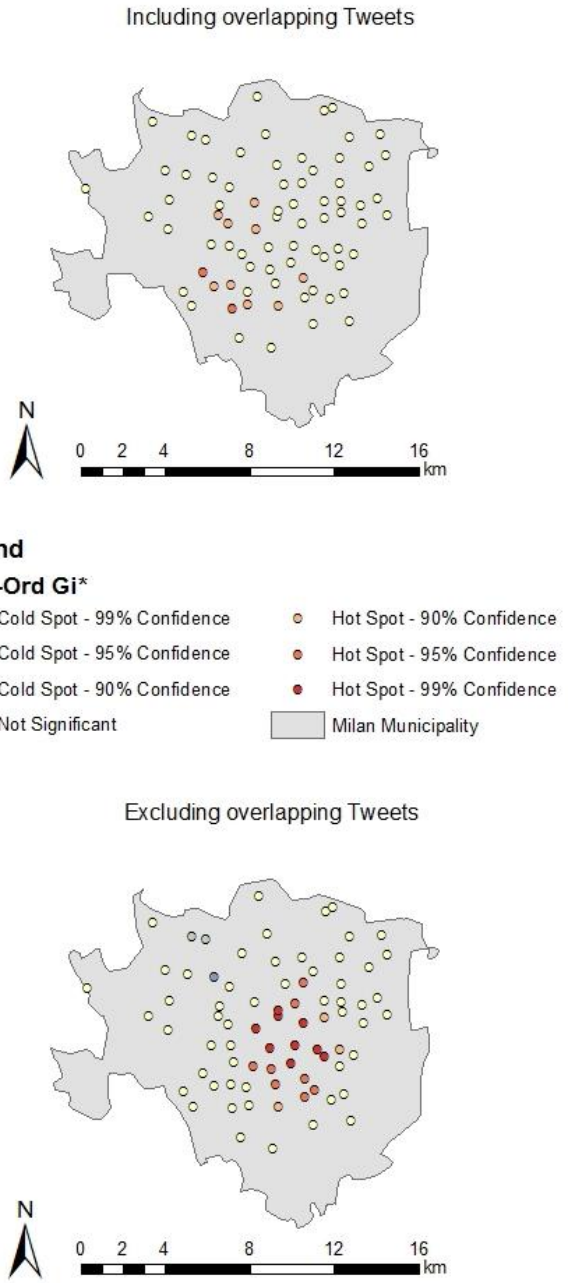

Figure 5. Getis-Ord Gi* hot spot analysis of the Tweets falling into a 1 km x 1 km grid in the Milan Municipality area; in the top panel all Tweets are considered; in the bottom panel the overlapping Tweets are excluded.

As it is well known, metadata have a key role in appropriately deciding if the data are "fit for use". In the case of Tweet geolocation, when the position accuracy is not known the spatial analyses that could be performed could turn out to be meaningless. Hopefully, in the next future these metadata could become available, enabling to take full advantage of social media data also from the spatial point of view.

ACKNOWLEDGMENTS

The authors wish to thank QuestFactory (Pavia, Italy) for providing the data used in this paper for research purposes and, in particular, Ms Irene Liberali (M.Sc.) for her useful explanations.

REFERENCES

[1] Steiger, E, Westerholt, R and Zipf, A. Research on social media feeds – A GIScience perspective. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F, Purves, R. (eds.) European Handbook of Crowdsourced Geographic Information, Pp. 237–254. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.r. License: CC-BY 4.0., 2016

[2] Roick, O., Heuser, S., Location based social networks–definition, current state of the art and research Agenda. Transactions in GIS 17.5, pp. 763-784, 2013

[3] Di Martino, S., Romano, S., Bertolotto, M., Kanhabua, N., Mazzeo, A., Nejdl, W., Towards Exploiting Social Networks for Detecting Epidemic Outbreaks. Global Journal of Flexible Systems Management: 1-11, 2017

[4] Gordon, J., Comparative geospatial analysis of Twitter sentiment data during the 2008 and 2012 US Presidential elections, 2013

[5] Earle, P.S., Bowden, D.C., and Guy M., Twitter earthquake detection: earthquake monitoring in a social world. Annals of Geophysics 54.6, 2012

[6] Sakaki, T, Makoto O., and Yutaka M., Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on World wide web. ACM, 2010

[7] Leetaru, K, Wang, S., Cao, G., Padmanabhan, A., Shook, E., Mapping the global Twitter heartbeat: The geography of Twitter. First Monday 18.5, 2013

[8] Hahmann, S., Purves, R.S., Burghardt, D., Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. Journal of Spatial Information Science 2014.9,pp. 1-36, 2014

[9] Heravi, B. R., Ihab S., Tweet location detection. Computation+ Journalism Symposium, 2015

[10] Maximilian, W., Kaisser, M., Geo-spatial event detection in the twitter stream. European Conference on Information Retrieval. Springer Berlin Heidelberg, 2013

[11] Andrienko, G., Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., Thom, D. Thematic patterns in georeferenced tweets through space-time visual analytics. Computing in Science & Engineering 15.3, pp. 72-82, 2013

[12] Getis, A., and K. Ord., The Analysis of Spatial Association by Use of Distance Statistics. Geo- graphical Analysis 24, pp. 189-206, 1992